BEYOND THE LOCAL VIEW: PREDICTING CORRECTION GAINS FOR SPARSE ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation guides a student on its own prefixes, but local uncertainty and teacher–student disagreement do not directly measure the benefit of a correction. We define correction gain as the change in eventual success probability when a teacher candidate replaces the student's sampled token at a fixed prefix, with the current student continuing both branches. Correction-Gain Selection (CGS) learns candidate values from offline paired interventions and uses their predicted difference to select 20% of response positions for distillation. The predictor remains frozen while its inputs follow the evolving student, and training retains the original trajectories. Across five mathematical benchmarks, CGS improves macro-average accuracy over the strongest matched-budget sparse baseline by 1.73 and 1.69 percentage points for Qwen3-1.7B and DeepSeek-R1-Distill-Qwen-1.5B, and by 1.81 points for Qwen3-8B to Qwen3-4B-Base distillation. Paired diagnostics in the two smaller-student settings connect this allocation to correction gains and examine predictor reuse as the student changes. The approach turns correction experience into reusable guidance about where teacher supervision helps the current student.

1 INTRODUCTION

On-policy distillation (OPD) guides students on prefixes they generate (Lin et al., 2020; Agarwal et al., 2024; Gu et al., 2024), addressing the distribution shift induced by a learner's own decisions (Ross et al., 2011; Bengio et al., 2015). Reasoning traces expose intermediate choices (Wei et al., 2022), and repeated sampling reveals alternative solution paths (Wang et al., 2023). Along these paths, teacher–student disagreement mixes consequential decisions with wording preferences and equivalent expressions. Improving supervision quality requires identifying guidance that helps the student complete its solution.

A *local view* evaluates positions through next-token distributions. Entropy measures student uncertainty; Local-KL measures teacher–student disagreement and, here, is also the distillation loss. These signals inform token importance and local teachability (Xu et al., 2026; Wang et al., 2026), but neither directly measures a suggestion's effect on eventual success. An uncertain wording choice may leave an answer unchanged, while a confident formula choice may derail the derivation.

Verification research distinguishes feedback on completed answers from feedback on intermediate reasoning (Cobbe et al., 2021; Uesato et al., 2022; Lightman et al., 2024). For teacher guidance, the student's continuation is central: the student must carry a proposed correction through the remaining reasoning. Its value therefore depends on the prefix, the suggestion, and the continuing student. Success after a teacher token alone leaves out what that student could achieve after its original choice.

We introduce Correction-Gain Selection (CGS): *value a teacher suggestion by the success it adds for the current student*. At a fixed prefix, CGS compares eventual success probabilities after the teacher's highest-probability token and the student's sampled token, with the same student continuing both branches. Their difference ranks positions by predicted correction gain. This criterion ties supervision to what the current student can gain from teacher guidance; benchmark evaluation measures the effect of training with that allocation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

A small candidate-value predictor learns from offline paired interventions across questions, positions, and checkpoints. Its parameters remain frozen during OPD, while its inputs follow the evolving student. Two scores per position allocate 20% supervision on the original trajectory, turning correction experience into a reusable selection criterion.

Our contributions are threefold: a student-relative criterion for incremental teacher help; a lightweight predictor that implements this criterion for sparse OPD; and benchmark results in three teacher–student settings, supported by paired correction diagnostics and checkpoint analyses in the two smaller-student settings.

2 Related Work

Learning on student-generated trajectories. Knowledge distillation transfers teacher predictions (Hinton et al., 2015); sequence-level distillation uses teacher-generated outputs (Kim & Rush, 2016). Supervision on learner-visited states addresses sequential distribution shift (Ross et al., 2011), as developed for autoregressive distillation by ImitKD and GKD (Lin et al., 2020; Agarwal et al., 2024). CGS studies supervision allocation within these trajectories.

Distillation objectives and data distributions. MiniLLM develops reverse-KL distillation (Gu et al., 2024); f-DISTILL studies sequence-level divergences and their step-wise decomposition (Wen et al., 2023). DistiLLM combines skew-KL with adaptive off-policy sampling, and DistiLLM-2 contrasts teacher- and student-generated responses (Ko et al., 2024; 2025). CGS varies selected positions while holding the loss fixed.

Selecting and adapting OPD supervision. TIP studies uncertainty and divergence; TA-OPD evaluates local teachability through teacher mass on the student's candidate support (Xu et al., 2026; Wang et al., 2026). Other approaches address thinking-pattern compatibility, prefix drift, and objectives adapted to teacher entropy (Li et al., 2026; Yang et al., 2026; Jin et al., 2026). SPOT probes continuations to calibrate targets (Qu et al., 2026). CGS learns candidate values offline; a frozen predictor uses their same-prefix value difference to select positions for the common reverse-KL loss without online candidate continuations.

Learning from reasoning outcomes. Outcomes support solution verification (Cobbe et al., 2021), rationale self-training (Zelikman et al., 2022; Singh et al., 2024), and mathematical reinforcement learning (Shao et al., 2024). Process supervision labels intermediate steps (Uesato et al., 2022; Lightman et al., 2024); Math-Shepherd automates labeling, and ReST-MCTS* couples value estimation with search (Wang et al., 2024; Zhang et al., 2024). CGS uses offline outcomes to learn supervision allocation.

Outcome information at intermediate decisions. Critical Tokens Matter identifies failure-associated tokens (Lin et al., 2025); Process Q-value Models rank step quality (Li & Li, 2025); process advantage verifiers measure progress through changes in continuation success under a prover policy (Setlur et al., 2025). CGS compares two tokens at one prefix under student continuation, using the original student choice as the baseline. The reference measures what the teacher adds to the student's own alternative. This comparison controls distillation locations while preserving the teacher-distribution target.

3 CGS: Supervision through Incremental Teacher Help

3.1 A student-relative value of guidance

Continuation outcomes provide a way to evaluate intermediate decisions, as in automatic process supervision and process advantage verification (Wang et al., 2024; Setlur et al., 2025). We specialize this perspective to two candidate tokens at the same student prefix. Let π_S and π_T denote the student and teacher. For a generated response of length L, let p_{t-1} include the prompt and response prefix, a_t^S be the token actually sampled by the student, and $a_t^T = \arg\max_a \pi_T(a \mid p_{t-1})$. Define candidate value by

$$V_{\pi_S}(p_{t-1}, a) = \Pr(\text{correct final answer} \mid p_{t-1}, a;\ \text{continuation by } \pi_S)\,. \tag{1}$$

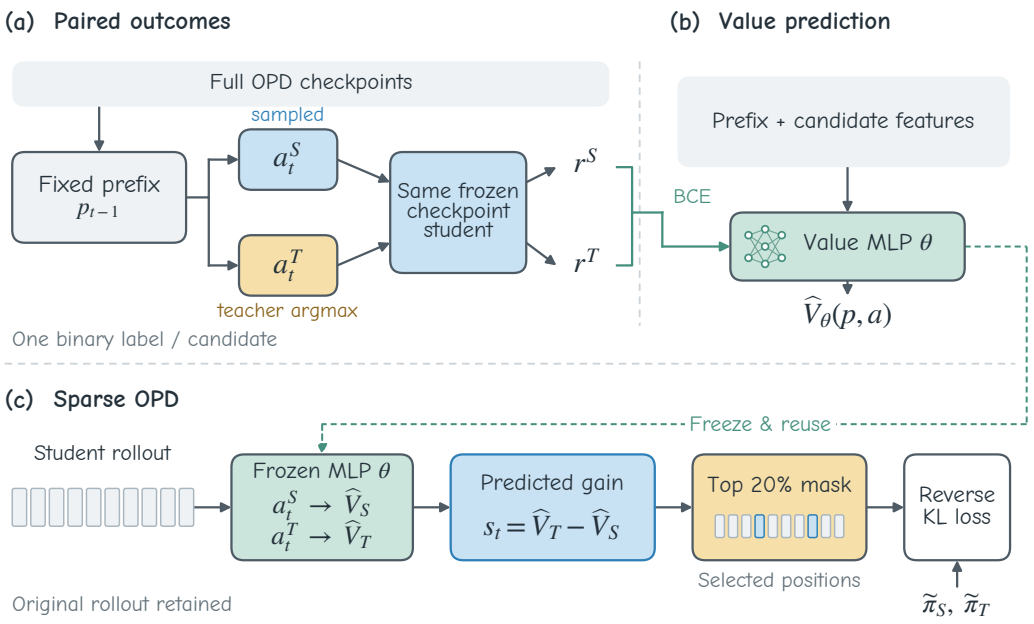

Figure 1: **CGS overview.** (a) Full OPD checkpoints supply prefixes; the original student token and teacher argmax token are each followed by the same checkpoint student to obtain binary outcomes. (b) One predictor scores both candidates from prefix and candidate representations plus four scalar features. (c) During OPD, the frozen predictor ranks positions by teacher-minus-student value and selects $\lceil 0.2L \rceil$ positions for reverse-KL supervision on the original trajectory.

This definition holds the prefix fixed and changes only the candidate before the student generates the suffix. A candidate-value predictor $\widehat{V}_\theta$ scores both alternatives, giving the position score

$$s_t = \widehat{V}_\theta(p_{t-1}, a_t^T) - \widehat{V}_\theta(p_{t-1}, a_t^S). \tag{2}$$

The original student candidate supplies a matched reference for the teacher's suggestion. Identical candidates have zero score. A high score expresses a predicted correction gain under the student's continuation, rather than the teacher candidate's success probability alone.

3.2 Learning reusable candidate values

For each setting, 2,000 questions and Full OPD checkpoints at rounds 0, 112, and 225 provide 6,000 source responses (Figure 1). We uniformly sample four distinct-candidate positions per response and obtain one binary outcome per candidate by continuing with the corresponding frozen student: 48,000 labels, or 24,000 pairs. A question-level split assigns 1,800 questions to training and 200 to validation, keeping all stages and candidates of a question together (43,200/4,800 labels).

Each of the three settings, including Qwen3-8B to Qwen3-4B-Base, trains its own candidate-value MLP with this protocol; that predictor scores both candidates. It combines the student prefix representation, the candidate's student embedding, and four scalars: both models' candidate log-probabilities and entropies. Projected vectors, their elementwise product, and projected scalars feed a small MLP. The Qwen3-1.7B and DeepSeek predictors have approximately 1.26M and 1.0M parameters. Appendix C.1 gives the architecture and fitting configuration.

We fit unweighted binary cross-entropy, a strictly proper probability loss (Gneiting & Raftery, 2007), with AdamW (Loshchilov & Hutter, 2019); validation NLL selects the checkpoint. Individual continuations provide binary labels, and the predictor learns probabilities across questions, positions, and stages. During OPD, its weights remain frozen and dropout is disabled while student-derived features follow the current student. Appendix G accounts for offline labels and a direct-continuation alternative.

3.3 APPLYING THE CRITERION TO SPARSE DISTILLATION

For each response, CGS scores all valid response positions and selects the largest $\lceil 0.2L \rceil$ scores. Valid positions include EOS and exclude prompt, template, and padding tokens. Ties are resolved by Local-KL and then position order. The budget is filled according to this ranking even when fewer than $\lceil 0.2L \rceil$ positions have positive scores.

Following reverse-KL language-model distillation (Gu et al., 2024), all selection methods use the same truncated objective. Define

$$\mathcal{A}_t = \mathrm{Top16}(\pi_S(\cdot \mid p_{t-1})) \cup \{a_t^T\}, \qquad \widetilde{\pi}_j(a \mid p_{t-1}) = \frac{\pi_j(a \mid p_{t-1})}{\sum_{b \in \mathcal{A}_t} \pi_j(b \mid p_{t-1})}, \tag{3}$$

for $a \in \mathcal{A}_t$ and $j \in \{S, T\}$. If M is the collection of selected positions in the training batch, the loss is

$$\mathcal{L}_{\mathrm{OPD}} = \frac{1}{|M|} \sum_{t \in M} D_{\mathrm{KL}}(\widetilde{\pi}_S(\cdot \mid p_{t-1}) \,\|\, \widetilde{\pi}_T(\cdot \mid p_{t-1})) \,. \tag{4}$$

Local-KL ranks positions by the same per-position divergence. Teacher candidates and predictor features come from the existing model forwards; the selector adds two small-MLP evaluations per position. Candidate replacement is used to obtain offline labels and diagnostic outcomes. Formal OPD trains on the unmodified student trajectory.

4 EXPERIMENTS

We evaluate trained-student performance, the selection criterion, and predictor reuse.

4.1 EXPERIMENTAL SETTING

Models and data. Table 1 covers three model pairs: Qwen3-1.7B with teacher `Keven16/Qwen3-4B-Non-Thinking-RL-Math-Step500` (thinking disabled) (Yang et al., 2025; Keven16, n.d.); DeepSeek-R1-Distill-Qwen-1.5B with JustRL-DeepSeek-1.5B (DeepSeek-AI et al., 2025; He et al., 2025); and Qwen3-4B-Base with Qwen3-8B (Yang et al., 2025). Teachers remain frozen. The diagnostics use the first two pairs, abbreviated Qwen and DeepSeek below. Base denotes the student before OPD. The MATH training set (Hendrycks et al., 2021) is split by question into 500 diagnostic, 2,000 labeling, and at least 3,600 OPD training questions. Separate MATH500 test responses (Lightman et al., 2024) support position statistics and the case study.

Training and comparison. Each of 225 rounds samples nine questions and eight responses per question, yielding 72 responses per round and 16,200 planned per run. We use full-parameter FSDP (Zhao et al., 2023) and AdamW at a constant learning rate of 10^{-6}. Full OPD supervises all valid positions; Random, Entropy, Local-KL, and CGS select 20% per response using uniform sampling, student uncertainty, local divergence, and predicted correction gain, respectively. Appendix C gives sampling, optimization, and response-budget details.

Evaluation and paired diagnostics. We evaluate AIME24, AIME25, AMC23, MATH500, and HMMT25 with 16 responses per question, temperature 0.7, top-p 0.95 (Holtzman et al., 2020), and a 31,744-token response cap. Avg@16 is accuracy over all sampled answers; Avg weights the five datasets equally. The last boxed answer is checked for mathematical equivalence; incorrect, unboxed, or truncated responses fail. Intervention diagnostics allocate eight outcome slots per candidate, with suffix budgets, terminal handling, and paired bootstrap procedures specified in Appendix C.

4.2 DOES HELP-BASED SELECTION IMPROVE THE TRAINED STUDENT?

With 20% supervision, CGS improves macro-average accuracy over the strongest sparse baseline by 1.73, 1.69, and 1.81 percentage points (pp) for Qwen3-1.7B, DeepSeek-R1-Distill-Qwen-1.5B, and Qwen3-4B-Base, respectively (Table 1). Gains over Full OPD are 3.49, 3.83, and 2.64 pp.

Table 1: **Final mathematical performance (Avg@16).** Accuracy %; Avg equally weights the five datasets. Denominators are 480/480/640/8,000/480 responses from 30/30/40/500/30 questions. Bold marks block-wise best values, including ties; sparse methods use 20% supervision.

Method	AIME24	AIME25	AMC23	MATH500	HMMT25	Avg
Qwen3-1.7B						
Base	11.25	11.67	47.50	73.20	6.67	30.06
Full OPD	37.08	33.33	80.63	89.29	19.17	51.90
Random	41.04	33.33	78.59	**92.11**	17.50	52.52
Entropy	41.88	32.92	80.78	91.84	20.83	53.65
Local-KL	41.67	34.58	83.59	89.50	18.96	53.66
CGS	**42.92**	**35.63**	**85.16**	92.01	**21.25**	**55.39**
DeepSeek-R1-Distill-Qwen-1.5B						
Base	27.08	23.75	70.63	83.80	15.83	44.22
Full OPD	36.25	31.67	74.06	84.90	17.92	48.96
Random	39.17	27.50	76.09	84.80	18.75	49.26
Entropy	41.46	31.88	76.88	**86.34**	18.96	51.10
Local-KL	40.63	29.38	79.84	85.41	19.58	50.97
CGS	**42.08**	**33.54**	**81.72**	85.78	**20.83**	**52.79**
Qwen3-8B → Qwen3-4B-Base						
Base	9.17	6.67	43.13	66.00	3.33	25.66
Full OPD	10.63	10.21	45.47	73.10	5.63	29.01
Random	10.21	11.46	47.34	72.34	3.96	29.06
Entropy	10.63	11.67	46.72	73.65	5.83	29.70
Local-KL	11.46	11.25	47.50	72.93	6.04	29.84
CGS	**13.33**	**13.33**	**50.00**	**74.90**	**6.67**	**31.65**

Table 2: **Top-1 correction on failed trajectories.** Repair is % (successes) over 800 outcome slots per branch: 100 trajectories × eight slots. Gain subtracts student from teacher repair; 95% CIs use paired trajectory bootstrap.

Model	Selection	Student repair	Teacher repair	Gain, pp [95% CI]
Qwen	Random	2.75 (22)	3.75 (30)	1.00 [−0.75, 2.63]
	Entropy	3.38 (27)	4.63 (37)	1.25 [−0.75, 3.38]
	Local-KL	4.88 (39)	7.88 (63)	3.00 [0.63, 5.38]
	CGS	3.75 (30)	7.75 (62)	4.00 [1.63, 6.50]
DeepSeek	Random	5.63 (45)	5.50 (44)	−0.13 [−2.50, 2.38]
	Entropy	7.38 (59)	9.25 (74)	1.88 [−1.38, 5.00]
	Local-KL	6.13 (49)	10.75 (86)	4.63 [1.75, 7.38]
	CGS	4.63 (37)	12.88 (103)	8.25 [5.25, 11.25]

CGS leads on four of five datasets for each smaller student and all five for Qwen3-4B-Base. On MATH500, Random leads Qwen3-1.7B (92.11% versus 92.01%), and Entropy leads DeepSeek (86.34% versus 85.78%).

4.3 What does an incremental criterion capture?

For each of 100 failed trajectories per diagnostic model, we compare student and teacher candidates at one distinct-candidate position with the prefix and student continuation policy fixed. Random samples uniformly; the other methods choose their highest-ranked position.

The student reference changes the comparison (Table 2). On Qwen, CGS and Local-KL yield 62 and 63 teacher-branch successes, but gains of 4.00 and 3.00 pp; DeepSeek gains are 8.25 and 4.63 pp. The paired CGS-minus-Local-KL differences are 1.000 pp [−2.125, 4.375] and 3.625 pp [−0.250, 7.375]. Thus teacher success and incremental help can rank selections differently.

Local ranks do not show a common monotonic relation with correction value (Figure 2). Qwen Entropy yields 5.328 pp [0.314, 10.307] in the 20–40% rank bin and 3.466 pp [2.502, 4.395] in the

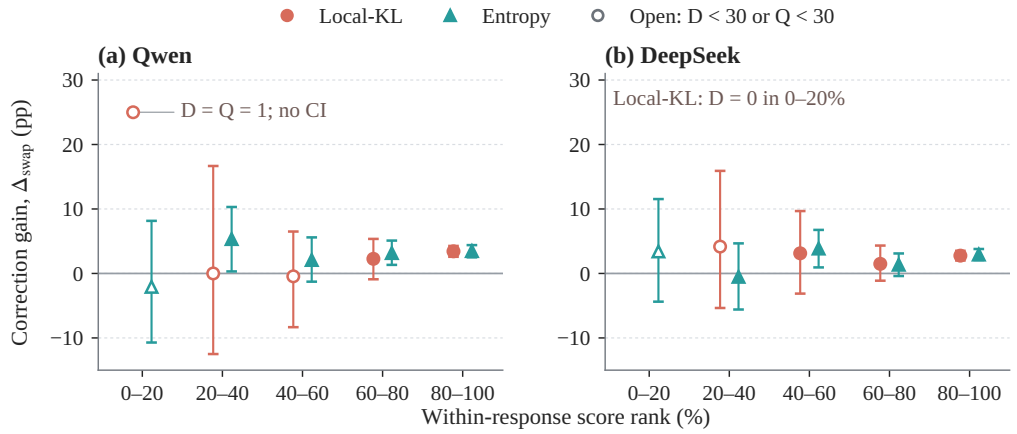

Figure 2: **Local-score ranks and correction gains.** An independent 500-question initial-student panel per model includes successful and failed responses. Sampling four positions from each Random, Entropy, and Local-KL mask gives 6,000 records and 5,972 unique positions, binned by within-response score rank. Gain is $100(T - S)/(8D)$; D counts distinct-candidate positions and S, T successful student/teacher outcome slots. Bars: 95% question-bootstrap CIs (2,000 resamples). Hollow, unconnected markers have $D < 30$ or fewer than 30 contributing questions. The lowest Local-KL bin has one distinct-candidate Qwen position (25 pp, no CI) and no distinct-candidate DeepSeek positions (missing). Appendix D provides counts and interval rules.

Table 3: **Overlap with CGS supervision.** Mean ± SD of per-question Jaccard on 500 diagnostic responses per model; each mask selects $\lceil 0.2L \rceil$ valid positions.

Comparator	Qwen	DeepSeek
Random	0.1109 ± 0.0146	0.1109 ± 0.0138
Entropy	0.3606 ± 0.0273	0.3455 ± 0.0284
Local-KL	0.5262 ± 0.0615	0.4878 ± 0.0587

highest bin. High Local-KL positions still provide positive gains: 3.431 pp [2.585, 4.267] for Qwen and 2.762 pp [2.008, 3.513] for DeepSeek.

4.4 How does supervision move?

On the 500-question initial-student panel, Table 3 reports mean and SD of per-question Jaccard overlap, $|M_i \cap K_i|/|M_i \cup K_i|$, between full 20% masks. Partial overlap with Local-KL reflects both score relationships and Local-KL tie-breaking among zero-score positions.

In the full masks, CGS selects positive- or zero-score positions, with zero-score positions filling 34.11% of the Qwen budget and 23.70% of the DeepSeek budget (Figure 3c,d; Appendix B).

Observed corrections on failed trajectories vary by model (Figure 3a,b). DeepSeek gains are 2.389 pp [1.238, 3.444] for CGS-only positions and 1.401 pp [0.290, 2.589] for KL-only; Qwen gains are 0.691 pp [−0.421, 1.686] and 0.858 pp [−0.406, 2.150]. Including shared positions gives CGS/Local-KL gains of 2.230/1.848 pp for DeepSeek and 0.825/0.865 pp for Qwen. CGS has higher subset gains on DeepSeek, while Local-KL is slightly higher on Qwen (Appendix E).

4.5 Can one predictor remain useful as the student changes?

On 4,800 validation candidates per model, success AUC is 0.7494 [0.7305, 0.7688] for Qwen and 0.7533 [0.7310, 0.7752] for DeepSeek. Among 775/733 same-prefix pairs with discordant outcomes, direction AUC is 0.6403 [0.6029, 0.6770]/0.6310 [0.5897, 0.6695], with winner accuracy 61.42%/60.03% (ties count as one half).

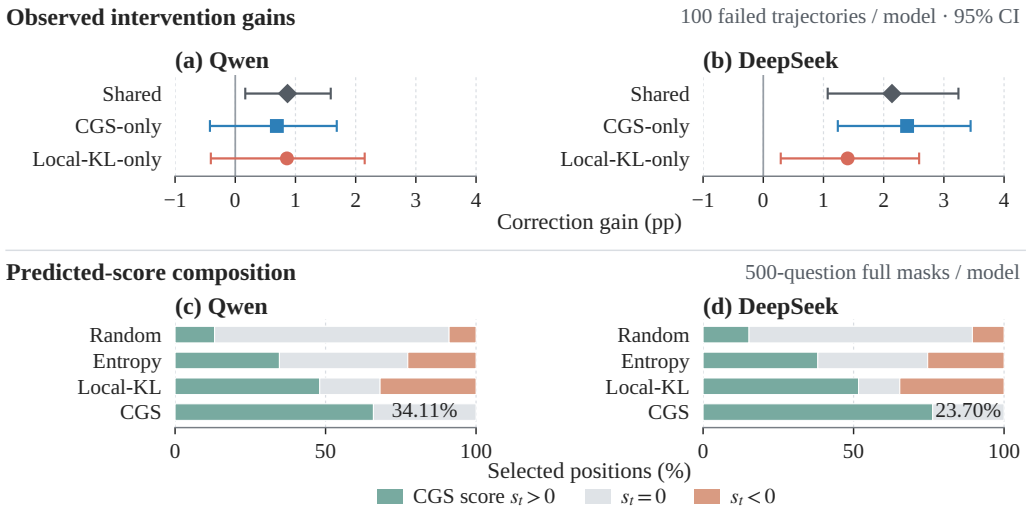

Figure 3: **Supervision composition and correction outcomes.** (a,b) On 100 failed trajectories per model, sample at most 12 distinct-candidate positions uniformly from the union of CGS and Local-KL masks. Shared/exclusive subsets reuse these outcomes; gains are $100(T-S)/(8N)$ with 95% CIs. Position counts (shared/CGS-only/KL-only): Qwen 677/217/306, DeepSeek 538/314/348. (c,d) Separately, full masks on 500 initial-student responses are partitioned by predicted gain sign: 185,215 selected positions per method for Qwen and 219,845 for DeepSeek.

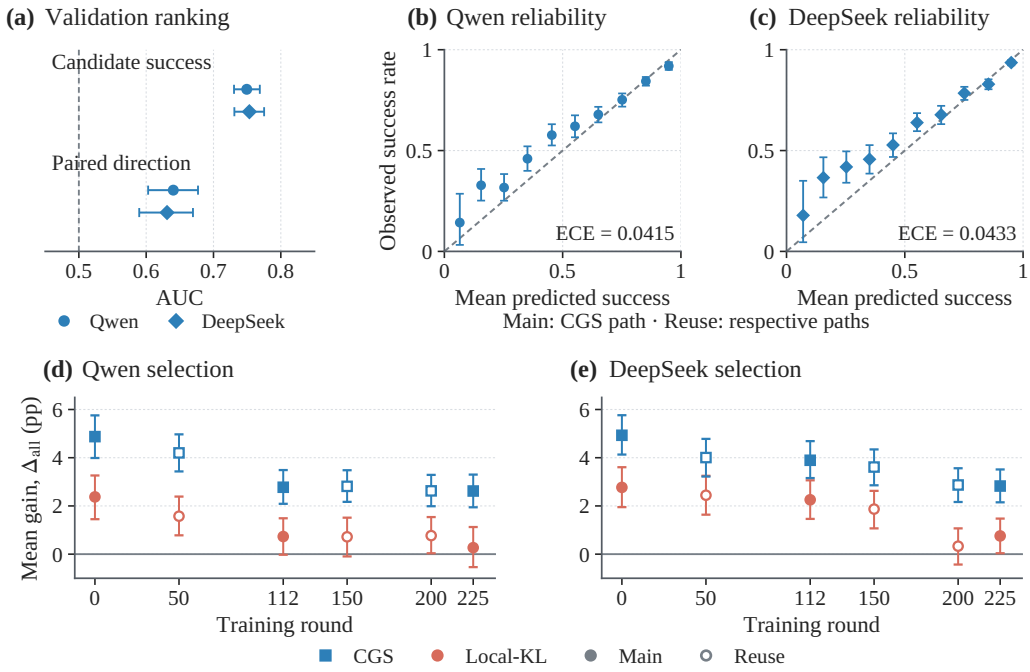

Figure 4: **Candidate judgment and continued use.** (a) Success and direction AUC on 4,800 validation candidates from 200 questions per model; direction uses 775/733 discordant pairs. (b,c) Calibration in 10 equal-width probability bins. (d,e) All-position gain Δ_{all} (Eq. 5). Filled points compare both masks at common CGS-path checkpoints (rounds 0/112/225); open points use each method's own training path (50/150/200). Each method samples 2,000 positions from 500 questions per checkpoint. Bars: 95% question-bootstrap CIs (2,000 resamples). Predictors trained on Full OPD rounds 0/112/225 remain frozen; reuse is within each model and task setting.

Calibration analysis (Guo et al., 2017) gives pooled ECE of 0.0415 [0.0338, 0.0551] for Qwen and 0.0433 [0.0316, 0.0561] for DeepSeek (Figure 4b,c). Teacher-candidate ECE exceeds student-candidate ECE (0.0666/0.0309 and 0.0767/0.0216). Appendix A gives stage metrics and calibration bins.

At each checkpoint, we sample four positions per mask from each of 500 questions ($N = 2{,}000$). Let D count distinct-candidate positions and S, T their successful student/teacher outcome slots. We distinguish gain per distinct-candidate intervention from gain over the full sampled mask:

$$\Delta_{\text{swap}} = 100\frac{T-S}{8D}, \qquad \Delta_{\text{all}} = 100\frac{T-S}{8N} = \Delta_{\text{swap}}\frac{D}{N}. \tag{5}$$

Identical candidates contribute zero net gain to Δ_{all}. These stage diagnostics use a 7,168-token cap for both models; Appendix F reports all counts and both denominators.

On common CGS-path checkpoints, correction gains generally decrease over training. At round 225, Δ_{all} is 2.619 pp [1.944, 3.300] for Qwen CGS versus 0.269 pp [−0.538, 1.125] for Local-KL, and 2.825 pp [2.150, 3.513] versus 0.756 pp [0.038, 1.475] for DeepSeek.

For reuse on each method's own training path, CGS retains higher gains at rounds 50, 150, and 200. At round 200, paired differences in Δ_{all} are 1.856 pp [0.894, 2.794] for Qwen and 2.538 pp [1.606, 3.481] for DeepSeek. Bootstrap resampling preserves matched questions across methods and stages (Appendix F).

4.6 A DECISION THAT CHANGES THE CONTINUATION

In a 1,080-token Qwen response to a MATH500 parallel-line distance problem, the student mistakes the magnitude of the parallel projection for the perpendicular distance (Figure 5). Three independent fixed-prefix interventions illustrate which local decisions alter that reasoning.

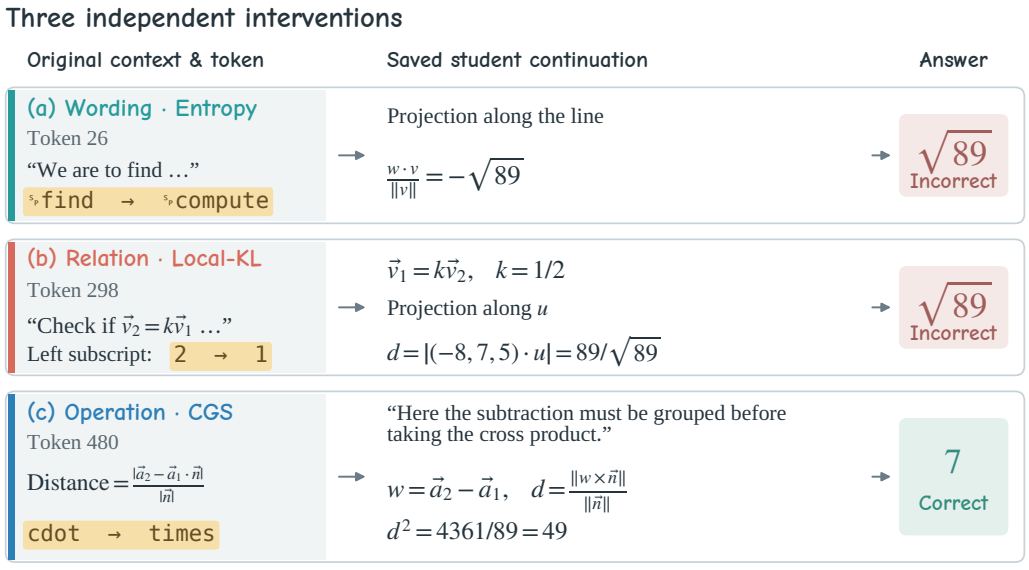

Figure 5: **A local choice and its downstream consequence.** Independent interventions at tokens 26/298/480 of Qwen288 (1,080 tokens, including EOS) preserve the original prefix. (a) A wording change and (b) an equivalent direction relation retain the wrong answer $\sqrt{89}$. (c) Changing the operation to a cross product yields $w \times n = (31, -6, 58)$ and $d = 7$; the continuation explicitly groups $w = a_2 - a_1$. The visible-space symbol marks a token-leading space.

Entropy's wording change (token 26) and Local-KL's subscript change (298) preserve the erroneous projection. CGS changes `cdot` to `times` (480); the student then groups the displacement and computes the cross product, obtaining $d^2 = 4361/89 = 49$ and the correct answer 7. The useful intervention changes a formula choice whose consequences the student can carry through.

In the separate 500-response MATH500 position analysis, CGS density peaks in the middle fifth for Qwen (23.84%) and DeepSeek (23.00%). Density divides selected tokens by all valid tokens in each segment, aggregated across responses (Appendix E).

5 Conclusion

CGS selects supervision by the teacher's predicted incremental help to the student. Offline paired outcomes train a candidate-value predictor that preserves the original OPD trajectory and loss while reducing supervised positions to 20%. Across three mathematical reasoning model pairs, CGS improves macro-average accuracy; diagnostics on two pairs connect selection to correction value and support frozen predictor reuse within each setting.

AI Use Statement

Generative AI tools (ChatGPT/Codex) assisted with manuscript drafting and revision, translation, literature search and citation organization, interpretation of reported results, and preparation of plotting and typesetting code. The authors are responsible for the final text, claims, references, and artifacts.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5be69a584901a26c521c2b51e40a4c20-Abstract-Conference.html.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/e995f98d56967d946471af29d7bf99f1-Abstract.html.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948v1.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. doi: 10.1016/j.patrec.2005.10.010. URL https://www.sciencedirect.com/science/article/pii/S016786550500303X.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8ac015d409635f196f9e3e9dcfb9a94e-Abstract-Conference.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Bingxiang He, Zekai Qu, Zeyuan Liu, Yinghao Chen, Yuxin Zuo, Cheng Qian, Kaiyan Zhang, Weize Chen, Chaojun Xiao, Ganqu Cui, Ning Ding, and Zhiyuan Liu. JustRL: Scaling a 1.5B LLM with a simple RL recipe. *arXiv preprint arXiv:2512.16649*, 2025. URL https://arxiv.org/abs/2512.16649.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874*, 2021. URL https://arxiv.org/abs/2103.03874. NeurIPS 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026. URL https://arxiv.org/abs/2603.07079v3. ICML 2026.

Keven16. Qwen3-4B-Non-Thinking-RL-Math-Step500. Hugging Face model repository, n.d. URL https://huggingface.co/Keven16/Qwen3-4B-Non-Thinking-RL-Math-Step500. Accessed September 26, 2026.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327. Association for Computational Linguistics, 2016. URL https://aclanthology.org/D16-1139/.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895. PMLR, 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. DistiLLM-2: A contrastive approach boosts the distillation of LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 31044–31062. PMLR, 2025. URL https://proceedings.mlr.press/v267/ko25a.html.

Wendi Li and Yixuan Li. Process reward model with Q-value rankings. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/26494b66ae1114d314673e25b4967288-Abstract-Conference.html.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026. URL https://arxiv.org/abs/2604.13016.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Alexander Lin, Jeremy Wohlwend, Howard Chen, and Tao Lei. Autoregressive knowledge distillation through imitation learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6121–6133. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.494. URL https://aclanthology.org/2020.emnlp-main.494/.

Zicheng Lin, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. Critical tokens matter: Token-level contrastive estimation enhances LLM's reasoning capability. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37906–37918. PMLR, 2025. URL https://proceedings.mlr.press/v267/lin25j.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Zikun Qu, Min Zhang, Mingze Kong, Zhiwei Shang, Zhengyu Chen, Yikun Ban, Shuang Qiu, and Zhongxiang Dai. SPOT: Sparse probing and outcome calibration for on-policy distillation. *arXiv preprint arXiv:2608.04419*, 2026. URL https://arxiv.org/abs/2608.04419v2.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635. PMLR, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/98711dea460bdefe0e651ca23ec98ba2-Abstract-Conference.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J. Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron Parisi, Abhishek Kumar, Alex Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshitiej Mahajan, Laura Culp, Lechao Xiao, Maxwell L. Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yundi Qian, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. Beyond human data: Scaling self-training for problem-solving with language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=lNAyUngGFK.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL https://www.jmlr.org/papers/v15/srivastava14a.html.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022. URL https://arxiv.org/abs/2211.14275.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in on-policy distillation. *arXiv preprint arXiv:2605.26844*, 2026. URL https://arxiv.org/abs/2605.26844.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. f-divergence minimization for sequence-level knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10817–10834. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.605. URL https://aclanthology.org/2023.acl-long.605/.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. TIP: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026. URL https://arxiv.org/abs/2604.14084v4.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Zhicheng Yang, Zhijiang Guo, Yifan Song, Minrui Xu, Yongxin Wang, Yiwei Wang, Xiaodan Liang, and Jing Tang. Prune-OPD: Efficient and reliable on-policy distillation for long-horizon reasoning. *arXiv preprint arXiv:2605.07804*, 2026. URL https://arxiv.org/abs/2605.07804.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/639a9a172c044fbb64175b5fad42e9a5-Abstract-Conference.html.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. ReST-MCTS*: LLM self-training via process reward guided tree search. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/76ec4dc30e9faaf0e4b6093eaa377218-Abstract-Conference.html.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. PyTorch FSDP: Experiences on scaling fully sharded data parallel. *Proceedings of the VLDB Endowment*, 16(12): 3848–3860, 2023. doi: 10.14778/3611540.3611569. URL https://www.vldb.org/pvldb/vol16/p3848-huang.pdf.

A PREDICTOR PERFORMANCE AND CALIBRATION

Each label-source checkpoint uses the same 200 validation questions, with four positions and two candidates per question. This gives 1,600 candidate labels per checkpoint and 4,800 labels pooled across checkpoints. Table A1 reports candidate-success AUC, expected calibration error (ECE), Brier score, and negative log-likelihood (NLL). AUC measures success ranking (Fawcett, 2006). ECE (Guo et al., 2017) uses 10 equal-width probability bins and weights each bin's absolute difference between observed success and mean prediction by its candidate count. Brier score (Brier, 1950) is the mean squared error of the predicted probability against the binary outcome, and NLL is the mean negative log-likelihood (Gneiting & Raftery, 2007); lower values indicate better probability estimates for both metrics.

For the pooled same-prefix comparison, 775 Qwen pairs and 733 DeepSeek pairs have discordant outcomes. The predictor chooses the successful candidate on 476/775 and 440/733 pairs, respectively: winner accuracies of 61.42% and 60.03%, with tied predictions counted as one half.

Table A1: **Predictor metrics by label-source checkpoint.** Successes count positive binary candidate labels. Pooled rows combine all three checkpoints.

Checkpoint	Candidates	Successes	AUC	ECE	Brier	NLL
Qwen						
Round 0	1600	1097	0.7499	0.0546	0.1806	0.5383
Round 112	1600	1185	0.7393	0.0425	0.1670	0.5061
Round 225	1600	1181	0.7523	0.0452	0.1667	0.5090
Pooled	4800	3463	0.7494	0.0415	0.1715	0.5178
DeepSeek						
Round 0	1600	1172	0.7576	0.0564	0.1652	0.5037
Round 112	1600	1218	0.7438	0.0464	0.1590	0.4881
Round 225	1600	1265	0.7561	0.0333	0.1475	0.4527
Pooled	4800	3655	0.7533	0.0433	0.1572	0.4815

Table A2: **Stage ECE by candidate type.** Each checkpoint contains 1,600 candidates: 800 student candidates and 800 teacher candidates.

Checkpoint	All candidates	Student candidates	Teacher candidates
Qwen			
Round 0	0.0546	0.0494	0.0851
Round 112	0.0425	0.0256	0.0612
Round 225	0.0452	0.0267	0.0692
DeepSeek			
Round 0	0.0564	0.0452	0.0932
Round 112	0.0464	0.0240	0.0825
Round 225	0.0333	0.0304	0.0675

Table A2 separates stage ECE by candidate type. Tables A3 and A4 pool both candidate types and all three checkpoints into 10 calibration bins per model. The observed success rate is successes divided by candidates. The signed calibration difference is observed success minus mean prediction, so positive values indicate underestimation. Its count-weighted absolute value recovers pooled ECE, up to rounding of the displayed values. The intervals in these tables refer to the calibration difference; the vertical intervals in Figure 4 refer to the observed success rate.

Table A3: **Qwen calibration bins.** The 4,800 candidate labels are pooled across both candidate types and three label-source checkpoints. The last column reports the signed calibration difference and its 95% confidence interval.

Probability bin	Candidates	Successes	Mean prediction	Observed success	Observed − predicted [95% CI]
$[0.0, 0.1)$	35	5	0.0652	0.1429	$0.0777\ [-0.0322, 0.2216]$
$[0.1, 0.2)$	125	41	0.1553	0.3280	$0.1727\ [0.0977, 0.2535]$
$[0.2, 0.3)$	199	63	0.2524	0.3166	$0.0642\ [0.0004, 0.1314]$
$[0.3, 0.4)$	296	136	0.3513	0.4595	$0.1081\ [0.0487, 0.1690]$
$[0.4, 0.5)$	352	203	0.4542	0.5767	$0.1225\ [0.0709, 0.1753]$
$[0.5, 0.6)$	461	286	0.5522	0.6204	$0.0682\ [0.0144, 0.1220]$
$[0.6, 0.7)$	585	397	0.6508	0.6786	$0.0278\ [-0.0103, 0.0658]$
$[0.7, 0.8)$	721	542	0.7516	0.7517	$0.0002\ [-0.0339, 0.0319]$
$[0.8, 0.9)$	961	811	0.8519	0.8439	$-0.0080\ [-0.0298, 0.0127]$
$[0.9, 1.0]$	1065	979	0.9487	0.9192	$-0.0295\ [-0.0480, -0.0113]$

Table A4: **DeepSeek calibration bins.** The 4,800 candidate labels are pooled across both candidate types and three label-source checkpoints. The last column reports the signed calibration difference and its 95% confidence interval.

Probability bin	Candidates	Successes	Mean prediction	Observed success	Observed − predicted [95% CI]
$[0.0, 0.1)$	28	5	0.0702	0.1786	$0.1084\ [-0.0248, 0.2763]$
$[0.1, 0.2)$	93	34	0.1558	0.3656	$0.2098\ [0.1115, 0.3113]$
$[0.2, 0.3)$	160	67	0.2528	0.4188	$0.1659\ [0.0897, 0.2420]$
$[0.3, 0.4)$	208	95	0.3509	0.4567	$0.1058\ [0.0336, 0.1764]$
$[0.4, 0.5)$	273	144	0.4501	0.5275	$0.0774\ [0.0174, 0.1352]$
$[0.5, 0.6)$	404	258	0.5517	0.6386	$0.0869\ [0.0453, 0.1329]$
$[0.6, 0.7)$	471	319	0.6537	0.6773	$0.0236\ [-0.0231, 0.0674]$
$[0.7, 0.8)$	747	586	0.7517	0.7845	$0.0328\ [-0.0005, 0.0644]$
$[0.8, 0.9)$	1071	888	0.8537	0.8291	$-0.0245\ [-0.0499, -0.0004]$
$[0.9, 1.0]$	1345	1259	0.9499	0.9361	$-0.0139\ [-0.0269, -0.0017]$

B MASK COMPOSITION AT A FIXED 20% BUDGET

Table A5 reports the full masks for the initial student on the independent 500-question diagnostic panel used for mask overlap and the composition analysis in Figure 3. For every selection method, selected positions are classified by the sign of the CGS predicted teacher-minus-student value difference s_t. Positive, zero, and negative therefore describe predicted candidate-value differences.

Table A5: **Predicted-score composition of the selected masks.** All methods use a 20% per-response budget. Sign counts use the same CGS score s_t for every method.

Method	Selected positions	$s_t > 0$	$s_t = 0$	$s_t < 0$
Qwen				
Random	185215	24324	144299	16592
Entropy	185215	64199	79016	42000
Local-KL	185215	89029	37131	59055
CGS	185215	122045	63170	0
DeepSeek				
Random	219845	33545	163158	23142
Entropy	219845	83796	80265	55784
Local-KL	219845	113558	30135	76152
CGS	219845	167731	52114	0

In these masks, CGS fills each $\lceil 0.2L_i \rceil$ budget with positive-score positions followed by zero-score positions ranked by Local-KL. Zero-score selections contribute to the overlap; total budgets sum the per-response integer budgets.

C Experimental and statistical protocols

C.1 Candidate labels and predictor fitting

Each setting uses 2,000 labeling questions at Full OPD checkpoints 0, 112, and 225, with one source response per question and checkpoint. Uniformly sampling four distinct-candidate positions per response and two candidates per position gives 6,000 responses, 24,000 paired positions, and 48,000 binary candidate labels. A question-level split assigns 1,800 questions (43,200 records) to training and 200 (4,800 records) to validation, keeping all stages and candidates of each question together.

The predictor uses the student prefix hidden state $h(p)$, the student's candidate embedding $e(a)$, and four scalars: student and teacher candidate log-probabilities and entropies. The vectors are projected to 256 dimensions each and the scalars to 32 dimensions. Concatenating the two vector projections, their elementwise product, and the scalar projection yields 800 dimensions, followed by an $800 \to 256 \to 1$ probability head. The Qwen3-1.7B and DeepSeek-R1-Distill-Qwen-1.5B predictors have approximately 1.26M and 1.0M parameters, respectively. Both candidates share one predictor within a setting; settings use separate predictors.

Fitting uses unweighted binary cross-entropy, AdamW with learning rate 0.001 and weight decay 0.01, batch size 512, and dropout 0.1 (Srivastava et al., 2014). Training runs for at most 50 epochs, with validation-NLL early stopping and patience five. During OPD, predictor weights are frozen, dropout is disabled, and inputs follow the current student.

C.2 Formal OPD configuration and model scope

The formal training schedule has 225 rounds, each sampling nine questions and eight responses per question: 72 responses per round and a planned 16,200 responses per run. Full-parameter FSDP uses FP32 master parameters and BF16 forwards. AdamW has constant learning rate 10^{-6}, no warmup, and weight decay 0.01. Sampling uses temperature 1.0 and top-p 1.0; prompts are capped at 1,024 tokens. The 9×8 configuration specifies sampling, not a per-device micro-batch or optimizer-update count.

The Qwen diagnostic setting uses Qwen3-1.7B and teacher `Keven16/Qwen3-4B-Non-Thinking-RL-Math-Step500`, with thinking disabled; the DeepSeek setting uses DeepSeek-R1-Distill-Qwen-1.5B and JustRL-DeepSeek-1.5B. The additional main-table setting uses Qwen3-4B-Base as student and frozen Qwen3-8B as teacher, with a separately trained candidate-value MLP following the common predictor-fitting, distillation, and evaluation protocol. Candidate diagnostics, predictor metrics, mask analyses, and checkpoint reuse are reported for the first two settings. The Qwen3-4B-Base setting follows the Qwen3-1.7B training configuration and runs on four NVIDIA A800 GPUs with 80 GB of memory each. The response-cap names Qwen and DeepSeek below refer to the two diagnostic settings.

C.3 Response budgets and outcome evaluation

Response lengths include EOS and exclude prompts, templates, and padding. Labeling, formal training, and the separate MATH500 position analysis use caps of 7,168 tokens for Qwen and 16,384 for DeepSeek. Formal training for Qwen3-4B-Base also uses a 7,168-token response cap. The initial-student mask panel and stage diagnostics use 7,168 for both. Final evaluation uses a 31,744-token response cap, temperature 0.7, top-p 0.95, and 16 responses per question.

An intervention's suffix budget equals the applicable response cap minus its fixed response length, including the candidate. Diagnostics allocate eight outcome slots per candidate. Nonterminal branches with remaining budget sample independent suffixes; EOS is judged once and reused across slots. Non-EOS branches with exhausted budgets fail. The last boxed answer is checked for mathematical equivalence; incorrect, unboxed, and length-truncated responses fail.

Avg@16 counts correct responses among all sampled answers; the macro-average weights datasets equally. Repair rates condition on originally failed trajectories. Correction gain subtracts student-branch from teacher-branch success. Stage gains use distinct-candidate or all sampled positions

as denominators, with identical candidates contributing zero net difference, as defined in Eq. 5. Subgroup gains within the common evaluation union are weighted by position counts.

C.4 Resampling and sparse bins

The uncertainty estimates use nonparametric bootstrap resampling (Efron, 1979). Top-1 diagnostics bootstrap failed trajectories with both candidate branches together. Question-level analyses retain all positions and paired candidates of each sampled question; predictor analyses also retain its label-source checkpoints. Predictor and question-level diagnostic intervals use 2,000 bootstrap replicates, describing sampling uncertainty under their respective data constructions.

In Figure 2, D distinct-candidate positions from Q contributing questions yield $8D$ outcome slots per candidate and gain $100(T-S)/(8D)$ pp. Zero-denominator replicates are omitted from interval quantiles. No interval is reported for $Q < 2$ or fewer than 95% nonempty replicates. Bins with $D < 30$ or $Q < 30$ are marked as sparse; an observed $D = 0$ bin has missing gain. These within-response score ranks differ from response-position segments.

C.5 Mask budgets and response-position statistics

Each response of valid length L_i receives $\lceil 0.2L_i \rceil$ selected positions. Summing these integer budgets can produce an aggregate fraction slightly above 20%. CGS ranks predicted value differences, breaking ties by Local-KL and then position order. In Appendix B, positive-score positions precede zero-score positions when filling the budget.

Position statistics use 500 MATH500 responses per model. Selection precedes division at $\lfloor jL_i/5 \rfloor$, $j = 0, \ldots, 5$. With selected and valid-token counts M_{ij} and B_{ij}, segment density is $100 \sum_i M_{ij} / \sum_i B_{ij}$. Qwen CGS selects 90,929 of 453,625 tokens; DeepSeek selects 510,398 of 2,551,002. This panel and the complete-mask panel in Appendix B use separate responses, lengths, and masks.

D COMPLETE SCORE-RANK DIAGNOSTICS

The initial-student panel contains 500 questions per model, including successful and failed responses. Random, Entropy, and Local-KL each contribute four positions sampled from their full 20% masks per question. Deduplication by question and position reduces 6,000 records to 5,972 positions. Both score axes use this same union. Ranks are computed within each response, with 0–20% the lowest quintile; the union contributes unequal numbers of positions to the bins.

Table A6 gives the counts underlying Figure 2. Here N counts positions, D counts distinct-candidate positions, and Q counts questions contributing those positions. Student and teacher success counts S, T each use $8D$ outcome slots; identical-candidate positions are not continued. The gain is $100(T - S)/(8D)$ pp. Interval construction and missing-value rules follow Appendix C.4.

Table A6: **Complete score-rank bins.** Intervals are 95% question-bootstrap intervals. B_{valid} counts nonempty resamples out of 2,000. A dagger marks $D < 30$ or $Q < 30$.

Rank (%)	N	D	Q	S/T	Δ_{swap}, pp [95% CI]	B_{valid}
Qwen: Local-KL						
0–20†	416	1	1	6/8	25.000 (no CI)	1266
20–40†	455	5	5	22/22	0.000 [−12.500, 16.667]	1987
40–60†	548	28	28	127/126	−0.446 [−8.333, 6.501]	2000
60–80	866	182	153	951/984	2.266 [−0.904, 5.357]	2000
80–100	3687	2896	500	15191/15986	3.431 [2.585, 4.267]	2000
Qwen: Entropy						
0–20†	423	18	17	100/97	−2.083 [−10.714, 8.157]	2000
20–40	412	61	60	312/338	5.328 [0.314, 10.307]	2000
40–60	581	204	176	1076/1110	2.083 [−1.274, 5.593]	2000
60–80	915	499	342	2564/2690	3.156 [1.330, 5.097]	2000
80–100	3641	2330	498	12245/12891	3.466 [2.502, 4.395]	2000
DeepSeek: Local-KL						
0–20†	395	0	0	0/0	— ($D = 0$)	0
20–40†	463	12	12	57/61	4.167 [−5.357, 15.909]	2000
40–60	559	36	35	183/192	3.125 [−3.127, 9.678]	2000
60–80	891	235	198	1264/1292	1.489 [−1.124, 4.325]	2000
80–100	3664	3127	500	17285/17976	2.762 [2.008, 3.513]	2000
DeepSeek: Entropy						
0–20†	417	22	22	95/101	3.409 [−4.376, 11.538]	2000
20–40	427	71	66	424/421	−0.528 [−5.593, 4.665]	2000
40–60	560	201	174	1062/1124	3.856 [0.937, 6.754]	2000
60–80	858	492	331	2795/2848	1.347 [−0.398, 3.101]	2000
80–100	3710	2624	500	14413/15027	2.925 [2.082, 3.793]	2000

For Qwen's lowest Local-KL bin, the estimate comes from one distinct-candidate position in one question; 1,266 of 2,000 resamples are nonempty. The next bin has 1,987 nonempty resamples. DeepSeek's lowest Local-KL bin has no distinct-candidate positions, so its gain is missing. All other bins have 2,000 nonempty resamples.

E SUPERVISION SUBSETS AND RESPONSE POSITIONS

E.1 SHARED AND EXCLUSIVE SUPERVISION POSITIONS

The subset analysis in Figure 3 uses the same 100 failed trajectories per model as the top-1 diagnostic. For each response, the full CGS and Local-KL masks are unioned, distinct-candidate positions are retained, and at most 12 positions are sampled uniformly. Shared, CGS-only, and KL-only subsets inherit their outcomes from this common evaluation union. The evaluated CGS set combines Shared and CGS-only; the evaluated KL set combines Shared and KL-only. Each branch has $8N$ outcome slots, and aggregate gains weight subsets by their position counts.

Table A7: **Counts and correction gains in the common evaluation union.** Coverage is the number of represented failed trajectories. Gains are $100(T - S)/(8N)$ pp with 95% intervals.

Subset	N	Coverage	S	T	$8N$	Gain, pp [95% CI]
Qwen						
Shared	677	100	200	247	5416	0.868 [0.166, 1.586]
CGS only	217	88	41	53	1736	0.691 [−0.421, 1.686]
KL only	306	98	85	106	2448	0.858 [−0.406, 2.150]
CGS eval	894	100	241	300	7152	0.825 [0.246, 1.400]
KL eval	983	100	285	353	7864	0.865 [0.281, 1.462]
DeepSeek						
Shared	538	100	214	306	4304	2.138 [1.070, 3.242]
CGS only	314	97	84	144	2512	2.389 [1.238, 3.444]
KL only	348	98	118	157	2784	1.401 [0.290, 2.589]
CGS eval	852	100	298	450	6816	2.230 [1.449, 3.005]
KL eval	886	100	332	463	7088	1.848 [1.061, 2.634]

E.2 SELECTION DENSITY ALONG A RESPONSE

Table A8 reports all five relative-position segments for the separate 500-response MATH500 panel. Each method selects over the full response before segmentation. Percentages divide selected tokens by all valid tokens in the corresponding segment, aggregated across responses, as defined in Appendix C.5. Qwen uses a 7,168-token response cap and DeepSeek 16,384.

Table A8: **Selection density by response position.** All densities are percentages. The final column gives CGS selected-token and valid-token counts for each segment.

Segment (%)	Entropy	Local-KL	CGS	CGS selected / valid
Qwen				
0–20	22.84	24.49	17.20	15567/90521
20–40	21.12	20.75	21.72	19705/90732
40–60	18.72	17.79	23.84	21626/90711
60–80	19.61	18.39	21.63	19621/90732
80–100	17.95	18.82	15.85	14410/90929
DeepSeek				
0–20	12.09	20.35	15.89	81052/509992
20–40	17.84	20.24	20.30	103558/510213
40–60	22.61	20.39	23.00	117351/510186
60–80	27.71	21.36	22.53	114976/510213
80–100	19.79	17.71	18.31	93461/510398

E.3 Parallel-line intervention example

The Qwen288 case in Figure 5 is a 1,080-token MATH500 response, including EOS, about the distance between

$$(3,-10,1)+t(2,-9,-2), \qquad (-5,-3,6)+u(4,-18,-4).$$

With displacement $w=(-8,7,5)$ and direction $n=(2,-9,-2)$, the original solution uses the magnitude of the parallel projection, $|w\cdot n|/\|n\|=\sqrt{89}$. Independent interventions occur at 1-based tokens 26, 298, and 480, or 2.4%, 27.6%, and 44.4% of the response. Each preserves the original prefix. The CGS-selected token changes `cdot` to `times`; in the continuation, the student groups the full displacement $w=a_2-a_1$ and computes $w\times n=(31,-6,58)$, giving $d^2=4361/89=49$ and $d=7$.

F Training-stage diagnostics and frozen reuse

The stage diagnostics use a fixed panel of 500 questions per model, including trajectories whose original answers are correct and incorrect. For each method and question, four positions are sampled uniformly from its 20% mask, giving $N=2000$ positions per method. Of these, D have distinct student and teacher candidates. Each candidate at these positions has eight outcome slots; S and T count successful student- and teacher-candidate outcomes, respectively. Identical candidates contribute zero net gain. We report correction gains in percentage points (pp) using both denominators:

$$\Delta_{\text{swap}}=100\frac{T-S}{8D}, \qquad \Delta_{\text{all}}=100\frac{T-S}{8N}=\Delta_{\text{swap}}\frac{D}{N}.$$

These diagnostics use a 7,168-token response cap for both models. The label-generation and training caps are 7,168 for Qwen and 16,384 for DeepSeek. Confidence intervals use 2,000 bootstrap resamples clustered by question.

Table A9 evaluates both selection methods at fixed checkpoints from the CGS training path, at rounds 0, 112, and 225. Mean gains decrease across these checkpoints for both methods, while CGS retains higher gains than Local-KL under both denominators.

Table A9: **Correction gains across training stages.** Each model block uses the same 500-question panel and $N=2000$ sampled positions per method and checkpoint. Brackets give 95% confidence intervals.

Round	Method	D/N	S	T	Δ_{swap} (pp) [95% CI]	Δ_{all} (pp) [95% CI]
Qwen						
0	Local-KL	1564/2000	8252	8632	3.037 [1.866, 4.159]	2.375 [1.450, 3.263]
0	CGS	1261/2000	6474	7254	7.732 [6.374, 9.042]	4.875 [3.988, 5.756]
112	Local-KL	1458/2000	8567	8684	1.003 [−0.035, 2.021]	0.731 [−0.025, 1.488]
112	CGS	1053/2000	6215	6659	5.271 [3.991, 6.629]	2.775 [2.088, 3.488]
225	Local-KL	1482/2000	8906	8949	0.363 [−0.721, 1.512]	0.269 [−0.538, 1.125]
225	CGS	1054/2000	6198	6617	4.969 [3.703, 6.233]	2.619 [1.944, 3.300]
DeepSeek						
0	Local-KL	1686/2000	9353	9796	3.284 [2.316, 4.283]	2.769 [1.950, 3.606]
0	CGS	1467/2000	8032	8821	6.723 [5.660, 7.824]	4.931 [4.131, 5.763]
112	Local-KL	1668/2000	10007	10368	2.705 [1.757, 3.663]	2.256 [1.462, 3.063]
112	CGS	1452/2000	8601	9224	5.363 [4.347, 6.452]	3.894 [3.156, 4.688]
225	Local-KL	1579/2000	9657	9778	0.958 [0.048, 1.870]	0.756 [0.038, 1.475]
225	CGS	1253/2000	7759	8211	4.509 [3.448, 5.589]	2.825 [2.150, 3.513]

The predictor learns from Full OPD checkpoints at rounds 0, 112, and 225, then remains frozen during CGS training. Table A10 examines rounds 50, 150, and 200, using the respective checkpoints of the Local-KL and CGS training paths. Reuse is evaluated within the same model and task setting, with the same 500 questions, $N = 2000$ positions per method, and two gain denominators. Paired question-level bootstrap resampling retains all method and stage records for each sampled question; Table A11 reports the resulting method differences.

Table A10: **Frozen-predictor reuse across training rounds.** Counts and correction gains follow the stage-diagnostic protocol. Brackets give 95% confidence intervals.

Round	Method	D/N	S	T	Δ_{swap} (pp) [95% CI]	Δ_{all} (pp) [95% CI]
Qwen						
50	Local-KL	1565/2000	8361	8613	2.013 [1.006, 3.059]	1.575 [0.781, 2.388]
50	CGS	1237/2000	6468	7140	6.791 [5.584, 8.047]	4.200 [3.431, 4.969]
150	Local-KL	1478/2000	8784	8899	0.973 [−0.129, 2.032]	0.719 [−0.094, 1.513]
150	CGS	1096/2000	6373	6823	5.132 [4.030, 6.299]	2.813 [2.169, 3.481]
200	Local-KL	1519/2000	8510	8633	1.012 [0.048, 2.013]	0.769 [0.037, 1.538]
200	CGS	1109/2000	6074	6494	4.734 [3.592, 5.883]	2.625 [1.988, 3.288]
DeepSeek						
50	Local-KL	1729/2000	9810	10201	2.827 [1.892, 3.744]	2.444 [1.637, 3.250]
50	CGS	1537/2000	8577	9218	5.213 [4.200, 6.198]	4.006 [3.219, 4.781]
150	Local-KL	1670/2000	10212	10511	2.238 [1.281, 3.129]	1.869 [1.069, 2.625]
150	CGS	1436/2000	8682	9260	5.031 [4.004, 6.052]	3.613 [2.856, 4.344]
200	Local-KL	1651/2000	9951	10004	0.401 [−0.521, 1.297]	0.331 [−0.431, 1.069]
200	CGS	1295/2000	7707	8166	4.431 [3.376, 5.466]	2.869 [2.163, 3.563]

Table A11: **Paired differences in correction gain: CGS minus Local-KL.** Differences use matched questions at each round. All values are in pp, with paired-bootstrap 95% confidence intervals.

Round	Δ_{swap} difference [95% CI]	Δ_{all} difference [95% CI]
Qwen		
50	4.778 [3.297, 6.364]	2.625 [1.612, 3.706]
150	4.160 [2.655, 5.675]	2.094 [1.119, 3.119]
200	3.722 [2.241, 5.162]	1.856 [0.894, 2.794]
DeepSeek		
50	2.386 [1.114, 3.705]	1.563 [0.544, 2.650]
150	2.793 [1.529, 4.146]	1.744 [0.750, 2.769]
200	4.029 [2.758, 5.280]	2.538 [1.606, 3.481]

At all three reuse checkpoints, the paired method differences favor CGS under both gain denominators, with confidence intervals above zero.

G CANDIDATE-BRANCH ACCOUNTING

The value predictor separates an initial outcome-labeling investment from selection on subsequent trajectories. Table A12 compares extra candidate-branch budgets under a specified direct-continuation alternative: four positions per formal OPD response and one continuation for each candidate. The comparison counts branches, with no conversion to elapsed time or compute savings. The complete CGS workflow also includes obtaining Full OPD checkpoints, 6,000 source responses, predictor fitting and scoring, normal OPD rollouts and teacher forwards, diagnostics, and final evaluation.

Table A12: **Extra candidate-branch budgets per model.** The direct alternative evaluates four positions per OPD response, with two candidates and one continuation each. Counts are configured branch budgets, including terminal branches that can be judged directly. They do not measure generated tokens or runtime.

Scheme	Calculation	Candidate branches
Direct online continuations	$225 \times 72 \times 4 \times 2 \times 1$	129,600
CGS offline labels	$2{,}000 \times 3 \times 4 \times 2 \times 1$	48,000
Difference	$129{,}600 - 48{,}000$	81,600 (62.963%)
CGS extra continuations during OPD	Frozen predictor scoring	0